WOODHOUSE EXHIBIT 4

EXHIBIT F

Message

From: Xiaolan Wang [redacted]@meta.com]
Sent: 4/2/2024 11:03:32 PM
To: Nikolay Bashlykov [redacted]@meta.com]; David Esiobu [redacted]@meta.com]; Frank Zhang [redacted]@meta.com]; Xiaolan Wang [redacted]@meta.com]; Viktor Kerkez [redacted]@meta.com]
Subject: Message summary [{"otherUserFbId":null,"threadFbId":7614278578629298}]

David Esiobu (4/02/2024 10:08:20 PDT):
>oh @Xiaolan one thing i should have mentioned, if we're using fairspark for the download, we should be careful about disk use. i can probably help deploy separate machines for that so it doesn't crowd out other jobs running there
>
>if we use the GenAI AWS cluster then it should be fine, i think the fsx allocation is pretty large there

Xiaolan Wang (4/02/2024 10:19:56 PDT):
>Thanks for bring this up, yesterday, I was hitting some disk limitations on fairspark. Let's sync offline on this.

Nikolay Bashlykov (4/02/2024 10:20:04 PDT):
>for libgen I synced the files to S3 once a chunk was downloaded and removed them from cluster, so this can help save the space

Nikolay Bashlykov (4/02/2024 10:20:49 PDT):
>here is an example script:
https://github.com/fairinternal/[redacted]

David Esiobu (4/02/2024 12:48:12 PDT):
>@Frank can you clarify why we can't use FB infra for this again?

Frank Zhang (4/02/2024 15:39:55 PDT):
>avoiding risk of tracing back the seeder is from FB server

Frank Zhang (4/02/2024 15:40:15 PDT):
>avoiding risk of tracing back the seeder/downloader are from FB servers

Frank Zhang (4/02/2024 15:40:26 PDT):
>what's the argument for using FB infra?

Xiaolan Wang (4/02/2024 15:52:14 PDT):
>I think we are aligned on using aws for the downloading work. We were thinking whether we can get DI's support on the downloading part.
>
>Also, quick updates on data downloading:
>- Internet Archive: 65% documents are mirrored by AA and with torrents available. It has <200 torrents, but large in size (>2TB). The fastest way would be using a few aws nodes for the download. @David Esiobu will help find aws resources for us to process the download.
>- Z-library: 99% of the data are mirrored by AA with available torrents. Also has 200+ torrents, which we can use a few aws nodes for the download (similar to Internet Archive).
>- Libgen: torrents directly released by AA are not available for now. AA provided external torrents collection, which we can download from. The timestamp of these torrents are December 2023, which may potentially include extra docs compared to the previous downloads. There are 6000 torrents with much smaller chunks of data. We will use pyspark for the downloading work.
>- DuXiu: this is the hardest one to get with 0 torrents. I confirmed that we can still access the resource via URL links and direct download (of each individual document). Will be much slower to process and we will facing IP banned/browser verification problems to access all the data.
>
>Based on these, we can kick off Internet Archive/Z-library downloads when we get the AWS resource. Meanwhile, I will revise the pyspark code to re-download the Libgen. We will revisit DuXiu after kick off the other downloading jobs.

Frank Zhang (4/02/2024 16:01:43 PDT):
>sounds like a great plan!

Frank Zhang (4/02/2024 16:02:16 PDT):
>> get DI's support on the downloading
>the agreement is not involving DI at all, getting this effort in stealth mode

Frank Zhang (4/02/2024 16:03:32 PDT):
>there's an explicit decision to not involve DI. We want to get AA ingesting done in stealth mode within our group